日本語会話練習システムの開発に向けた音声の提示と評価に関する研究

博士学位論文要約（平成31年３月）

日本語会話練習システムの開発に向けた音声の提示と評価に関する研究

ハフィヤン　プラフィアント
指導教員：伊藤　彰則

Study on Presentation and Evaluation of Speech Toward Development of a Japanese Oral Communication Training System

Hafiyan Prafianto
Supervisor: Akinori ITO

Some computer-assisted language learning (CALL) systems have been made for Japanese language learners. The advance in speech technology has made listening practice and pronunciation training possible. In this paper, some ways to improve Japanese CALL systems were investigated. The effects of prosodic properties (speaking rate, pause position and pause length) on the perception of speech were investigated. Synthetic speech with various conditions was used to investigate how they correlate with the intelligibility and listenability of spoken Easy Japanese. The conditions for the ideal speaking speed was found. For pronunciation evaluation, a novel method to score the prosody of non-native speakers using parametric speech synthesis was proposed. The proposed method introduced feature substitution to reduce interference from the other pronunciation features. The result of the scoring experiment showed that the proposed method improved the scoring reliability in terms of the inter-rater correlation of the obtained scores. I also built automatic pronunciation evaluation systems to evaluate accent and compared the prediction performance using the conventional and the proposed scoring methods. The result showed that the predicted pronunciation scores by the proposed method were closer to human scores than that of the conventional scoring method in terms of correlation and prediction error.

1. Introduction

There are millions of foreign residents living in Japan nowadays, and the number of learners of the Japanese language outside of Japan is increasing as well. This means that the demand for Japanese language learning is very high. Recently, some systems of computer-assisted language learning (CALL) are being developed to meet this demand.

CALL systems offer various technologies[1)2)] to help learning languages at a lower cost than going to the language schools. For the Japanese language, CALL systems may offer kanji learning practice, vocabulary practice, and other functions related with reading and writing skills. The advances in speech technology has also made training for skills related with oral communication, namely listening and speaking, possible. Because Japan uses a writing system that is uncommon in other parts of the world, oral communication may be particularly important for the learners.

In this dissertation, I focused on two problems related with the CALL systems for Japanese oral communication skills, particularly the listening skill and the pronunciation skill. The effects of prosodic properties (speaking rate, pause position and pause length) on the perception of speech for listening training were investigated. Also, the use of automatic pronunciation evaluation for pronunciation training was considered. A novel method to score the prosody of non-native speakers using parametric speech synthesis was proposed to improve the reliability of the automatic evaluation.

2. Speech technologies for CALL

To improve the listening skill, a language learner must listen to a large amount of native utterance for various words and sentences. A CALL system can provide many utterances by using speech synthesis to generate sound simply from text, thus eliminating the need to record a large amount of native speech[3)]. For current speech synthesis technologies, many based on hidden Markov model[4)] (HMM), speech highly intelligible for non-native listeners can be made even though the speech might be less natural than a recorded speech.

To improve the pronunciation skill, a learner must practice pronouncing a large amount of utterances and get a feedback on how good their pronunciation is.

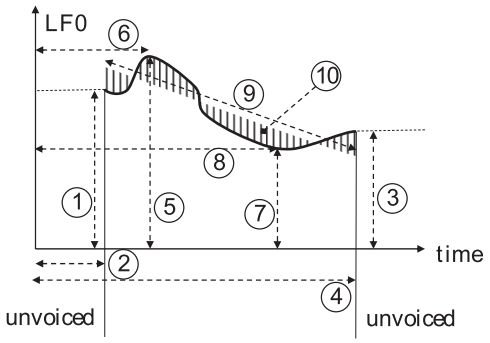

Fig. 1 The parameters used for the automatic evaluation of accent.

For pronunciation training, some CALL systems provide automatic pronunciation evaluations for the learners' utterances. Recent studies focus on using scores given by human raters to train machine learning models, e.g., support vector regression (SVR), clustering tree, and other methodologies[5)6)7)]. In this dissertation, the method based on SVR[6)] was used because it can be trained with few scores. First, the system extracted the following features that expressed the shape of the contour: (1) the onset of log F0 and (2) its position, (3) the offset of log F0 and (4) its position, (5) the maximum of log F0 and (6) its position, (7) the minimum of log F0 and (8) its position, (9) the slope of the regression line and (10) the total deviation from the regression line. These features are shown in Figure 1. Next, a model to predict accent scores was trained using SVR. After training, this model can then be used to predict the score of a new utterance by calculating the aforementioned features and using it as the input to the SVR model.

3. Presentation of speech with appropriate prosodic control

A CALL system can provide many utterances by using speech synthesis to generate sound simply from text, thus eliminating the need to record a large amount of native speech. Some systems generate the sounds with the speaking rate as fast as how a native speaker usually speak. However, for a beginner in learning, presenting a slower speaking rate might also be helpful. It is possible to control the rate of speaking in speech synthesis, and the function to play the sound at slower speeds is offered in a few CALL systems. However, there has only been very few studies on the best slow speed for non-native Japanese listeners.

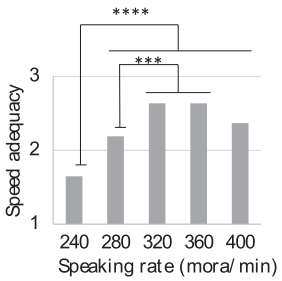

Fig. 2 Speed adequacy score for each speaking rate condition.

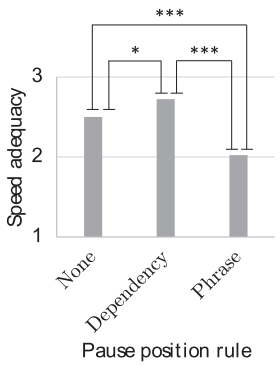

Fig. 3 Speed adequacy score for each pause position condition.

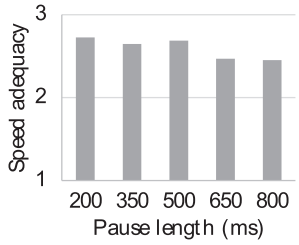

Fig. 4 Speed adequacy score for each pause length condition.

In this study, the speaking rate and pause properties that are preferred by Japanese language learners were investigated. A total of 67 subjects participated in the experiment. The subjects were asked to listen to synthesized speech with various conditions and rated how fast the sentence sound in a scale of 1 (very slow) through 3 (just right) to 5 (very fast). From this, the score of speed adequacy was calculated. Because

the option 3 is defined as “just right” for the perceived speed question, the perceived speed of 3 is defined as the most appropriate speed, with the speed adequacy of 3. The perceived speed of 2 and 4 have the speed adequacy of 2, while the perceived speed of 1 and 5 have the speed adequacy of 1.

The result shows that the speaking rate of 360 mora per minute (slower than the average native speaking rate at more than 470 morae per minute), with the pause inserted according to the rule based on dependency relation (same rule as native speakers), and pause length of 200 ms (similar as native speakers) were preferred. This is shown in Figures 2 to 4. Speech under these conditions was found to be more listenable and intelligible than speech at the standard speaking rate. Analysis of individual differences in the most adequate speaking rate showed a slight trend that participants with higher language proficiency prefer faster speaking rates. This suggests that as the learner's proficiency improves, the speaking rate can be made faster.

4. Evaluation of speech for specific prosodic properties

Some CALL systems provide automatic pronunciation evaluations for the learners’ utterances. Various methods have been proposed for this technology. Recently, machine-learning-based methods that relies on score given by native speakers has also been introduced. This technology uses a model trained using the acoustical features of non-native speakers' utterance as the input and the pronunciation scores of those utterances given by human native speakers as the output. The reliability of this method depends on the reliability of the scores given by human. For scores of specific prosodic properties such as accent and rhythm, the reliability is not very high. In this study, a novel method that utilizes parametric speech synthesis to improve the human scoring of accent and rhythm was proposed, and the effectiveness was investigated.

Here is the summary of the proposed method of scoring for accent. First, the scoring target, a non-native utterance, is recorded. Next, the parameter of the accent, the fundamental frequency (F0), is extracted. Then, speech synthesis based on the average voice model of native speakers is used to generate the parameters of the same utterance. Included in the generated parameters are the parameter of the accent, the parameter of the rhythm, and the parameter of the phoneme pronunciation.

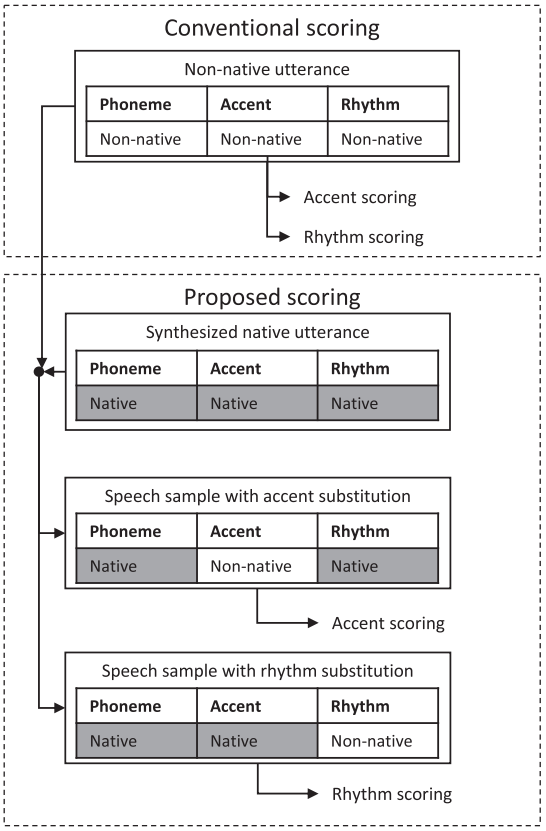

Fig. 5 Proposed scoring method using substitution of speech parameters

Next, the generated native accent parameter is substituted for the extracted non-native accent parameter. From those parameters, a speech sample whose accent is non-native but whose other features (phoneme pronunciation and rhythm) are native is generated. This new speech sample is then used in the scoring of accent. More reliable scoring can be expected for this speech sample because unlike the original non-native utterance, the features not being scored are native, and there will be less interference from those features to the scoring of accent. The score the rhythm, a similar procedure is done, but the non-native parameter of rhythm is substituted for the generated native rhythm.

The effectiveness of the proposed scoring method was investigated. To measure the reliability of the score, inter-rater correlation was investigated. To calculate this, seven raters were asked to rate the same four utterances, and the correlation between the scores were calculated.

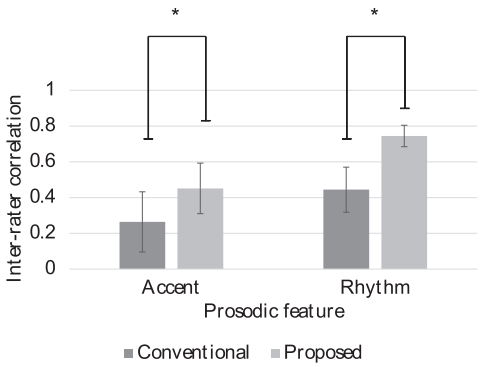

Fig. 6 Comparison of inter-rater correlations between the conventional and the proposed methods.

The experimental result, shown in Figure 6, shows that the proposed method can increase in the inter-rater correlation of accent and rhythm scores.

Finally, an automatic pronunciation evaluation system for accent was built with the utterance of ten non-native speakers speaking twelve Japanese words, with accent scores given by ten native Japanese raters using either the conventional scoring method and the proposed method. The result showed that the automatic evaluation model trained with the scores given by the proposed method could give scores closer to human scores than those of the conventional method.

5. Conclusions

In this dissertation, some ways to improve Japanese CALL systems were investigated. The effects of prosodic properties (speaking rate, pause position and pause length) on the perception of speech were investigated. Synthetic speech with various conditions was used to investigate how they correlate with the intelligibility and listenability of spoken Easy Japanese. The result showed that a speaking rate of 360 morae per minute with 200 ms pauses to be close to the ideal speaking speed. It is also more appropriate to insert pauses at appropriate natural positions for native speakers, based on the dependency relation rule of Japanese language, as opposed to inserting pauses between every phrases.

For pronunciation evaluation, a novel method to score the prosody of non-native speakers using parametric speech synthesis was proposed. The proposed method introduced feature substitution to reduce interference from the other pronunciation features. The result of the scoring experiment showed that the proposed method improved the scoring reliability in terms of the inter-rater correlation of the obtained scores. I also built automatic pronunciation evaluation systems to evaluate accent and compared the prediction performance using the conventional and the proposed scoring methods. The result showed that the predicted pronunciation scores by the proposed method were closer to human scores than that of the conventional scoring method in terms of correlation and prediction error.

References

1) M. Levy, “Technologies in use for second language learning,” The Modern Language Journal, vol.93, pp.769-782, 2009.
2) Y. Zhao, “Recent developments in technology and language learning: A literature review and meta-analysis,” CALICO journal, pp.7-27, 2003.
3) Z. Handley, “Is Text-To-Speech synthesis ready for use in Computer-Assisted Language Learning?” Speech Communication, vol.51, no.10, pp.906-919, 2009.
4) H. Zen, K. Tokuda, and A. Black, “Statistical Parametric Speech Synthesis,” Speech Communication, vol.51, no.11, pp.1039-1064, 2009.
5) M. Suzuki, T. Konno, A. Ito, and S. Makino, “Automatic Evaluation System of English Prosody Based on Word Importance Factor,” Journal of Systemics, Cybernetics and Informatics, vol.6, no.4, pp.83-90, 2008.
6) A. Maier, F. Hönig, V. Zeißler, A. Batliner, E. Körner, N. Yamanaka, P. Ackermann, and E. Nöth, “A Language-Independent Feature Set for the Automatic Evaluation of Prosody,” Proceedings of Interspeech, pp.600-603, 2009.
7) S.M. Witt, “Automatic Error Detection in Pronunciation Training: Where We Are and Where We Need to Go,” Proceedings of International Symposium on Automatic Detection of Errors in Pronunciation Training, pp.1-8, 2012.